Introduction and Motivation

While our original intent was focused around the association between music genre popularity and culture at large, we felt that its challenge was outside the scope of this project, and shifted gears towards what determines the popularity of songs

Data Description

We are using this data set, compiled by Michael Tauberg.

How the data was collected:

- Starting from the year 2000 onward, each week's archived Hot 100 was added to the data set, with their rank
- When a song is collected again, on a subsequent week, the original entry is removed with its rank updated and its previous value put into "last_pos"
- If the rank is higher than the previous week and the recorded peak position, the peak position is updated
- The song is then cross referenced with Spotify's API to acquire information about the song, such as characteristics, lyrics, and its spotify link.

While containing lots of useful information such as titles, genres, and lyrics, we really focused on the (spotify) musical characteristic combined with Peak Billboard Rank and year of release.

Mood: Danceability, Valence (Happiness), Energy, Tempo

Properties: Loudness, Speechiness, Instrumentalness

Context: Liveness, Acousticness

Segments: Key, Mode, Time Signature, Duration

Pre-Processing and Cleaning

Having loaded in the raw .csv with headers, converted it into a dataframe, and then took relevant columns and loaded them into a standalone data frame,w e made sure every column was numeric data, and as the data set had a significant amount of messing data, omitted incomplete rows.

Exploratory Data Analysis

We just plotted basic relationships between intuitive predictors and responses,

- Rank vs (Duration or Danceability or Energy)
- Year vs Danceability

Using the former, we discovered that the way that the data was collected -- removing the last time a song appeared from the data-set, to prevent duplicates, made "rank" unreliable, as songs would rarely exit the Hot 100 when they were still in the Top 40. Because of this, we opted for Peak Position, which was a lot more randomly distributed.

Once we accounted for this, we found clear -- albeit limited -- relationships between different musical characteristics, peak position, and year, and preceded with model selection.

Data Science Model

Our first model was a linear model. We used a cross validated LASSO approach to select our features.

```
cv.lasso <- cv.glmnet(x, y, family = "gaussian", alpha = .5, parallel = TRUE, standardize = TRUE,
type.measure = "mse")
coef(cv.lasso, s=cv.lasso$lambda.1se)
```

And then wrote the generalized linear model

```
fit <- glm(peak_pos ~ danceability + year + duration_ms + valence, data = df1)
```

As we saw room for improvement, we decided to form a polynomial regression model

```
poly <- glm(peak_pos ~ poly(energy, 2) + poly(liveness, 3) + poly(tempo,1)+ poly(acousticness,2) +
poly(instrumentalness,3) + poly(time_signature,3)+ poly(danceability,2) + poly(duration_ms,3)+
poly(loudness,1)+ poly(valence,1), data = df1)
```

We then converted that into a Generalized Additive Model:

```
gam <- gam(peak_pos ~ s(energy, 2) + s(liveness, 3) + s(tempo,1)+ s(acousticness,2) +
s(instrumentalness,3) + s(time_signature,3)+ s(danceability,2) + s(duration_ms,3)+ s(loudness,1)+
s(valence,1), data = df1)
```

And using backward stepwise feature selection, our GAM was simplified down to:

```
gam3 <- gam(peak_pos ~ s(liveness, 3) + s(tempo,1) + s(instrumentalness,3) + s(danceability,2) +
s(duration_ms,3)+ s(loudness,1)+ s(valence,1), data = df1) summary(gam3)
```

Model Evaluation

To evaluate the performance of our different models, we put them through 1016-fold cross validation to measure each Mean Standard Error

Model	MSE
Linear Model with all features	889.13
Linear Model with removed features	887.68
Polynomial Model with all features	885.83
GAM with all features	885.58
GAM with removed features	888.60

With this information, we deemed that while not very different; our most complex model (GAM with all features) had the most accurate predictive power of all the models.

Results

With our model, we learned that the musical qualities of a popular song is simply not an accurate indicator of where it will peak on the Hot 100 chart. Unsurprisingly, other characteristics may matter more -- such as artist popularity or marketing success among other things.

For perspective, here are some of 2018's #1 hits run through the linear model (which did not measure 2018):

Artist	Song	Predicted Peak Position
Drake	God's Plan	52
Ed Sheeran	Perfect	52
Childish Gambino	This is America	47
Cardi B	Bodak Yellow	48
Ed Sheeran	Shape of You	46

Two of these songs had a higher predicted peak position than the median and mean peak position of all of the songs in our data set.

Acknowledgements

Project conception: Dalvin Parks

Data: Billboard and Spotify

Compiled Data Set: Michael Tauberg

Models: Laura Setzer

Model evaluation and reworks: Dylan Mumm, Dalvin Parks, Laura Setzer

Visualizations: Dylan Mumm

Report: Dylan Mumm

Poster: Dalvin Parks

Software: R (libraries: boot, ggplot2, splines, glmnet, gam), Adobe Illustrator